April 2026

Police Handbook

for navigating cyber crimes

MOXII AFRICA

TABLE OF CONTENTS

ACRONYMS 4
DEFINITIONS 4
INTRODUCTION 6
WHAT ARE ONLINE HARMS? 8
- Online harassment 9
- Cyberstalking and cyberbullying 9
- Hate speech 10
- Incitement 11
- Non-consensual sharing of intimate images (NCII) and child sexual abuse material (CSAM) 11
- Cyber extortion 12
- Data breaches 13
- Financial cybercrime and fraud 13

EFFECTS OF ONLINE HARMS ON VICTIMS AND SURVIVORS 14
- The rights implications of online harms 14
- The psychological and social effects of online harms 16
- The physical and offline effects of online harms 16

AN OVERVIEW OF THE MAIN SOCIAL MEDIA PLATFORMS 18
- WhatsApp 18
- X (previously Twitter) 20
- Grok 22
- Instagram 23
- Facebook 25
- TikTok 29

EXISTING LEGISLATION 32
- Electronic Communications and Transactions Act 32
- The Cybercrimes Act 33
- The Protection from Harassment Act 34
- Prevention and Combating of Hate Crimes and Hate Speech Act 35
- The Criminal Law (Sexual Offences and Related Matters) Amendment Act 36

EMERGING TRENDS AND CASE STUDIES IN SOUTH AFRICA 37

PRACTICAL CONSIDERATIONS 38
Recognising patterns of online harm 38
Context is key 39
Challenges in addressing online harms 41
Collecting and preserving digital evidence 41
Investigating encrypted communications 41
Jurisdictional issues in cross-border harms 41
Cybersecurity and prevention strategies 42
Public awareness campaigns 42
Collaboration with tech platforms 43
Staying informed 43
CONCLUSION 44
ANNEXURE A: HANDLING COMPLAINANTS FOR ONLINE HARM: KEY STEPS 45
ANNEXURE B: EXAMPLE OF A POSTER THAT CAN BE PLACED WITHIN SAPS OFFICES 47
ANNEXURE C: INTERNET USER GUIDE FOR BEGINNERS 48

Acronyms

AI	Artificial Intelligence
CSAM	Child Sexual Abuse Material
DMs	Direct messages
ECTA	The Electronic Communications and Transactions Act 25 of 2002
ISC Africa	Internet Safety Campaign Africa
NCII	Non-Consensual distribution of Intimate Images
OSINT	Open-Source Intelligence
SAPS	South African Police Service
VPNs	Virtual Private Networks

Definitions

Topic	Definition
Online Harms:	Detrimental behaviours and practices occurring in the digital sphere that negatively affect individuals and communities.
Online Harassment:	Unwanted, aggressive behaviour, such as sending threatening or abusive messages, stalking, or spreading false information.
Misinformation and Disinformation:	Misinformation is false or misleading information shared without harmful intent; disinformation is deliberately false information to deceive. Both have the potential to cause harm by dividing society and spreading hate.
Data Protection and Privacy:	Measures to safeguard personal information from unauthorised access and misuse.
Content Moderation:	The process of monitoring and managing user-generated content to ensure compliance with platform rules and regulations.
Algorithmic Bias and Discrimination:	Unfair treatment of individuals or groups by algorithms due to biases in data or design

Extremism and Polarisation:	The amplification of extreme and divisive content can lead to social division and radicalisation.
Cybercrime and Fraud:	Illegal activities conducted online, such as hacking, phishing, and financial scams.
Psychological and Social Effects:	The impact of online harms on mental health and social relationships, such as anxiety, depression, and isolation.
WhatsApp:	A messaging app used for personal and group communication.
X (formerly Twitter):	A microblogging platform where users share short messages (tweets).
Instagram:	A photo and video-sharing platform.
Facebook:	A social networking site for sharing updates, photos, and links.
TikTok:	A video-sharing platform known for viral trends.
Emoticons/Emojis:	Small digital images or icons are used to express emotions, and ideas, or convey messages in text form.
Memes:	Humorous or cultural images, videos, or text that spread rapidly across the internet, often carrying a particular social or political message.
Retweeting/Resharing:	Sharing someone else's post or tweet with your own followers on platforms like X (formerly Twitter) amplifies the original message.
Reports:	Reporting is the action of notifying a platform about inappropriate content or behaviour that violates its guidelines or terms of service.
Stories:	Short, temporary posts on social media platforms (like Instagram or Facebook) that disappear after 24 hours.
Reels:	Short, engaging videos on platforms like Instagram and Facebook, designed to be fun and shareable.

1. Introduction

The internet has changed how we communicate, work, and connect, but with its benefits come serious risks. In South Africa and around the world, people are increasingly exposed to harmful behaviours online, otherwise known as online harms.

Online harms broadly refer to a range of behaviours and practices that occur on digital platforms and negatively affect individuals, communities, and society at large.[1]

These can include—

- online harassment;
- the sharing of sexual or violent content;
- hate speech;
- financial crimes;
- incitement of public violence
- extremist propaganda; and
- the abuse of private personal data.

In addition to causing damage to people's mental health, personal relationships, and sense of safety, many of these harms are also crimes. Moreover, online harms can directly lead to offline physical harms, both through physical harassment and/or violence, as well as through physical health complications that can be caused by anxiety and depression.[2] Vulnerable groups, especially women, children, foreign nationals, LGBTQIA+ individuals, and those already facing social exclusion, are often the most targeted.

1 World Economic Forum, 'Toolkit for Digital Safety Design Interventions and Innovations: Typology of Online Harms' (2023)(accessible **here**).

2 These can range from as headaches, fatigue and sleep-disturbances to self-harming practices and eating disorders. See NSPCC, 'Online abuse' (accessible **here**).

This handbook has been developed to assist members of the South African Police Service (SAPS) in understanding the different forms of online harm and how best to respond. It covers practical issues like how to recognise online harassment, identify digital privacy breaches, respond to illegal content such as the non-consensual distribution of intimate images (NCII) and child sexual abuse material (CSAM), and how to handle hate speech, extremist messages, and other online threats.

While existing cybercrime laws try to address these forms of content, interpretation and enforcement often vary, thus making it essential for members of SAPS to approach each case with care and an understanding of both the law and the rights of those involved. Technology can make online crimes hard to trace, but it also provides new tools to investigate them and support victims and survivors more effectively. This handbook aims to guide members of SAPS through these evolving challenges, equipping them to protect communities in the digital space with the same commitment they bring to the physical world.

2.
What are online harms?

Online harms are complex and often overlap, which makes them difficult to categorise neatly. They include various types of harmful behaviour, illegal content, and damaging online interactions. The consequence of some of these harmful behaviours is the violation of the individual's privacy and personal safety.

A helpful way to understand online harms is to look at three key dimensions: harm in content production, content distribution, and content consumption.

- **Production harms** may occur when people are physically harmed to create digital material, and this includes examples such as recording or livestreaming assault, rape, or child abuse.

- **Distribution harms** may arise when sensitive material (like intimate images) is shared online without consent. Even in instances where no one was harmed during production, the lack of consent then causes harm to the individual once the content is shared.

- **Consumption harms** occur when people are exposed to content that is illegal, misleading, or age-inappropriate. This would, for example, include content that incites violence against a person or a group of persons, damage to property, or other forms of illegal behaviour.

It is important to understand these overlapping categories. The nature and scale of harm will shape how a case is assessed and will better assist in determining the urgency of police intervention needed, how evidence should be processed and safeguarded, and the level of sensitivity needed when engaging with victims and survivors.

Disclaimer

The aim of this handbook is to provide members of SAPS with practical guidance and insight into the rising and evolving forms of online harms – specifically those that undermine personal dignity, safety, equality, and privacy. These harms often disproportionately affect women, children, foreign nationals and other communities at risk and require trauma-informed, survivor-centred policing responses.

While the handbook does not cover all forms of cybercrime in detail, it is important to recognise that many online harms intersect with broader cybercrimes such as phishing, malware, hacking, unauthorised access to devices or systems, and digital surveillance. These intersections may emerge during investigations and should be understood within a broader ecosystem of digital abuse and exploitation.

Whilst this is by no means a closed list, the following forms of online harm will be explored under this section:

- Online harassment, including cyberbullying, cyberstalking, hate speech, incitement, NCII, and CSAM;
- Cyber extortion and sextortion;
- Data breaches; and
- Financial cybercrimes.

Online harassment

Online harassment severely restricts individuals' ability to exercise their rights, particularly the rights to privacy, freedom of expression, and safety. This harassment manifests in various forms, including cyberbullying, cyberstalking, hate speech and incitement, and extortion. These behaviours aim to intimidate, silence, or shame their targets, often affecting marginalised groups like women, racial minorities, foreign nationals, and LGBTQ+ individuals.[3]

Cyberstalking and cyberbullying

Cyberbullying refers to repeated and hostile online behaviour aimed at harassing, intimidating, or shaming someone. It can occur across various platforms, including social media, messaging apps, gaming environments, and mobile phones.[4] Examples include sending abusive or threatening messages, impersonating someone via fake accounts, or using generative AI to target individuals with harassment.[5] Cyberbullying tends to be unpredictable and often driven by spite, and while common among children and teenagers, it can affect people of all ages.

3 United Nations Women, 'Online and ICT facilitated violence against women and girls during COVID-19' (2020)(accessible **here**).
4 UNICEF, 'Cyberbullying: What is it and how to stop it' (2025)(accessible **here**).
5 Id.

Cyberstalking, while closely related to cyberbullying, is generally more premeditated and targeted.[6] It typically involves an adult or experienced user who deliberately harasses, intimidates, or monitors a specific individual through digital platforms such as email, phones, social media, or message boards.[7] These actions are often motivated by control, anger, or revenge, and the psychological impact can be severe. Cyberstalking can take many forms, including excessive or threatening communication, the dissemination or threatened release of intimate images, impersonation through fake accounts and the incitement of third-party harassment.[8] In more serious cases, it may involve hacking to gain access to private information, steal sensitive data, impersonate the victim or survivor, or post malicious content.[9]

Both cyberbullying and cyberstalking demand prompt, trauma-informed responses from members of SAPS. The ability to assess severity, document patterns, and preserve digital evidence is critical in ensuring accountability and protecting survivors.

Hate speech

Hate speech is expression that goes beyond mere insults or offensive language – it infringes on the dignity of certain persons or groups.[10] Hate speech negatively impacts on victim or survivor's self-worth, and causes significant hurt, harm, pain, distress and humiliation.

Qwelane v South African Human Rights Commission and Another

While there is no universally accepted definition of the term "hate speech", the Constitutional Court outlined the following indicators in Qwelane v South African Human Rights Commission and Another to consider when determining whether expression amounts to hate speech:[11]

- Speech that goes beyond merely being offensive but also amounts to extreme hostility and vilification.
- Speech that can provoke discrimination against a group.
- Seech that targets people's rights based on their identity.

Notwithstanding these considerations, the Court highlighted that society should still tolerate different, unpopular or controversial views, noting that even disturbing or shocking opinions do not necessarily constitute hate speech.[12]

6 Social Media Victims Law Center, 'What is cyberstalking?' (2025) (accessible **here**).
7 Id.
8 Social Media Victims Law Center above n 6.
9 K Hanna, 'What is cyberstalking and how to prevent it?' TechTarget (2025) (accessible **here**).
10 South African Human Rights Commission, 'Hate Speech' (accessible **here**).
11 Qwelane v South African Human Rights Commission and Another [2021] ZACC 22; 2021 (6) SA 579 (CC); 2022 (2) BCLR 129 (CC) at para 81.
12 Id.

Incitement

Incitement occurs when a person encourages or urges others to commit acts of violence or hatred. When assessing incitement, it is not necessary that the action that was being incited actually has to occur.[13] Hate speech and incitement are frequently linked in that the effects are similar. Incitement may infringe on the targeted group or person's dignity and equality. However, incitement may also lead to physical harm.

Non-consensual sharing of intimate images (NCII) and child sexual abuse material (CSAM)

One of the most severe and emerging forms of online harassment is the NCII, which occurs when content depicting nudity, partial nudity, or sexually explicit acts is shared beyond its intended audience without consent.[14] It does not matter whether the person initially consented to the creation or limited sharing of the content; once disseminated without permission, it is NCII.[15] While people of all genders can be victimised, women are mostly affected. LGBTQIA+ persons and individuals with disabilities are also significantly affected.

Why words matter

Revenge porn versus NCII: Terminology such as NCII properly reflects the harm and removes the implied justification for the perpetrator's actions. Terminology is crucial in addressing online harms, particularly those that disproportionately affect marginalised groups. The term "revenge porn" suggests that the victim or survivor has committed an offense deserving of revenge, reinforcing victim-blaming narratives.

Child pornography versus CSAM: The term "child pornography" is misleading and harmful. It implies consent and even legality and draws harmful links to adult pornography, which is legal when it involves consenting adults. Children, however, cannot legally or ethically consent to sexual activity.

Using accurate, survivor-centred terminology like NCII and CSAM is not just semantics; it has real, practical consequences for how victims and survivors experience justice, support, and stigma. Terms like "revenge porn" and "child pornography" carry harmful implications. These labels can make victims and survivors feel complicit in their own victimisation, which could lead to feelings of shame and fear of being judged or disbelieved. As a result, many hesitate to report or seek help. By contrast, NCII and CSAM clearly name the abuse for what it is, which then helps survivors feel seen, validated, and more willing to come forward.

CSAM refers to photos, videos, or digital content that shows the sexual abuse or exploitation of a child. It is both a criminal offence and a grave violation of children's rights.[16] CSAM can be produced through coercion, deception, or manipulation, and its circulation continues to inflict

13 Southern Africa Litigation Centre, 'Hate Speech and Incitement' (accessible **here**).
14 Media Defence, 'Non-Consensual Sharing of Intimate Images' (accessible **here**).
15 Id.
16 Our Rescue, 'What is CSAM?' (2025) (accessible **here**).

trauma on victims and survivors long after the original abuse occurred. According to a household survey done in 2022, 9% of children reported being offered money or gifts in exchange for sexual images, another 9% were solicited to meet for sexual activities, and 7% had sexual images shared without their consent.[17] These figures may be under-reported due to the sensitive nature of the subject and children's reluctance to disclose abuse.

An alarming technological trend is the use of deepfake technology, where artificial intelligence (AI) manipulates images or videos to generate realistic but false content, usually for sexual exploitation.[18] The reach and impact of this content are intensified by the accessibility of open-source (accessible to the public) deepfake creation tools and widespread online communities dedicated to distribution.[19] With the sharing of deepfake NCII, two forms of harm are typically inflicted: first, upon the person whose face is manipulated into sexual content; and second, upon the original performer whose footage is misused.

Cyber extortion

Cyber extortion is a serious and growing form of digital crime in which an offender threatens an individual online unless or until the individual complies with the offender's demands.[20] Often, the offender would threaten to publish the targeted individual's sensitive data, expose private images, or launch cyberattacks until the demand is fulfilled. The extortionist typically gains access to valuable personal or corporate information, either by hacking into databases or by manipulating individuals into sharing private details, often through tactics such as impersonation or catfishing.[21] The techniques used vary depending on the target, with sophisticated attacks more often directed at organisations and more personalised approaches typically used to exploit individuals.[22]

Within this category, one particularly damaging tactic is 'sextortion', which is a form of sexualised blackmail wherein someone is coerced into sharing money, further images, or sexual acts under threat of having private, intimate material exposed.[23] These images may have been shared voluntarily, obtained without the targeted individual's knowledge, or fabricated altogether using AI-generated content.[24] Sextortion is frequently perpetrated through dating apps, live streaming platforms, or social media, where criminals pose as peers or trusted individuals before threatening to distribute compromising materials to family members, employers, or the public. While adults are often targeted, children are also increasingly being exploited through sextortion schemes. These acts leave young people feeling fearful, ashamed, and isolated – all of which could lead to severe emotional trauma and, in some tragic cases, suicide.[25]

---

17 ECPAT, INTERPOL, and UNICEF, 'Disrupting Harm in South Africa: Evidence of Online Child Sexual Exploitation and Abuse' (2022) (accessible **here**) at 7.
18 N Gotora, 'Unmasking deception: deepfake regulation in the context of South African law, could a rethinking of performers' protection rights be the answer?' (2024) International Journal of Law and Information Technology 32. (accessible **here**) at 6-7.
19 Id.
20 B Carpenter, 'Cyber Extortion' (29 August 2025) Digital Forensics (accessible **here**).
21 Id.
22 Id.
23 Internet Watch Foundation, 'Sextortion" or Sexually Coerced Extortion Help & Support' (accessible **here**).
24 Id.
25 Internet Watch Foundation, 'Exponential increase in cruelty as sextortion scams hit younger victims' (2024). (accessible **here**).

The consequences of online harassment often extend beyond the digital world. Victims and survivors may suffer severe psychological distress, fear for their physical safety, or face irreparable damage to their reputations.[26] In some cases, individuals may be forced to censor their views or withdraw from online spaces entirely. For instance, women who are disproportionately targeted by online sexualised harassment may feel compelled to self-censor, leaving important public conversations dominated by a less diverse group of voices.[27]

Data breaches

Social media platforms collect vast amounts of data from their users, often without the user's full understanding or consent. This collection includes not only the data users willingly provide, such as posts, photos, and location data, but also more covert data, such as interaction metrics. This includes what content users view, share, or engage with. As a consequence, social media platforms become frequent targets for cyberattacks due to the vast amounts of personal information they store. When data breaches occur, users' sensitive information can be exposed, leading to identity theft, financial fraud, or other malicious uses such as doxxing.

Doxxing explained

Doxxing refers to the malicious act of publicly revealing someone's private information without their permission, often with the intent to harass, intimidate, or harm. This can include real names, addresses, phone numbers, workplace details, and even behavioural patterns.

Because social media platforms have access to a lot of one's personal information, data breaches provide attackers with everything they need to either share personal information with the aim of "doxxing" victims and survivors or to carry out other malicious actions, such as those indicated above.

Financial cybercrime and fraud

Social media platforms have become fertile ground for various forms of financial cybercrime, including fraud, identity theft, phishing, and hacking. Cybercriminals exploit the openness and interconnectedness of social media platforms to target users with malicious schemes, often tricking them into sharing personal information or clicking on harmful links.[28] This poses significant risks to users' financial security and personal privacy. One particularly harmful form of cybercrime is online scams, where fraudsters pose as legitimate businesses, charities, or even friends to steal money or data.[29] These scams can range from phishing emails to complex schemes that exploit communities at risk, such as the elderly or those in financial distress.

26 Pew Research Center, 'The State of Online Harassment' (2021)(accessible **here**).
27 Id.
28 United Nations Office on Drugs and Crime, 'Cybercrime University Module Series' (accessible **here**).
29 Id.

3.

Effects of online harms on victims and survivors

Online harms affect victims and survivors in many ways. This section will describe these effects. First, the implications of online harms on victims and survivors' rights will be discussed. Further, the psychological, physical and social consequences of online harms will be set out. The offline manifestations of online harm will also be explained. Last, the particular effects of online harms on children will be discussed.

The rights implications of online harms

Online harms infringe on the constitutional rights of victims and survivors. The extent of infringement, as well as which rights are affected, is dependent on the type of harm as well as the degree of harm. Further, a victim or survivor's rights may be affected differently from another who has experienced the same type of harm, depending on that person's particular context. The concept of intersectionality is important to understand in this regard.

Intersectionality

Intersectionality recognises that individuals do not experience harm or discrimination through a single identity marker such as race, gender, or ability, but rather through the overlapping and compounding effects of multiple, interconnected characteristics.[30]

For example, the experiences of a Black transgender woman with a disability cannot be fully understood through the lens of gender or race alone; her exposure to online harm may be different from that of a white cisgender woman or a Black cisgender man. She may simultaneously face misogynistic abuse, racial slurs, transphobic attacks, and ableist hate.

Intersectionality should be kept in mind when assessing the impact of online harms on rights, as the infringements may manifest differently depending on a person's particular characteristics and circumstances.

Online harms may infringe on a variety of rights. The specific rights affected will differ from case to case, in the same way that the effects of a physical harm are context specific. The below listed rights are often infringed on by online harms. It is not a closed list.

- **Dignity:** The Constitution provides everyone with the right to have their human dignity respected and protected.[31] Online harms may infringe peoples' dignity by attacking their self-worth, privacy, or reputation, and by causing psychological distress. Certain online harms may humiliate, shame or stigmatise the victim or survivor, thereby infringing on their right to dignity.[32] The public nature of online platforms means that harmful content can be widely seen and shared, which may amplify this impact.

- **Privacy:** Depending on the nature of the harm in question, online harms may infringe on people's right to privacy.[33]

- **Freedom and security of the person:** Section 12 of the Constitution states that everyone has the right to be free from all forms of violence. Online harms may lead to offline violence, which also violates this right.

- **Equality:** It has been seen that online harms increasingly have gendered dimensions and disproportionately affect women through threats of sexual violence, misogynistic disinformation campaigns and NCDII.[34] Therefore, it may be that a woman victim or survivor's right to equality is infringed upon by an online harm.

30 T Meer and A Müller, 'Considering intersectionality in Africa' (2017)(accessible **here**) at 3.
31 Section 10 of the Constitution.
32 Report of the Special Rapporteur on Violence Against Women, Its Causes and Consequences on Online Violence Against Women and Girls from a Human Rights Perspective A/HRC/38/47 (2018)(accessible **here**) at para 57.
33 Section 14 of the Constitution.
34 African Commission on Human and Peoples' Rights, 'Resolution on the Protection of Women Against Digital Violence in Africa' ACHPR. Res. 522 (LXXII)(2022)(accessible **here**).

- **Children's rights:** Section 28 of the Constitution states that a child's best interests are of paramount importance in every matter concerning the child, and that they have the right to be protected from abuse or degradation. The best interests of the child are a dynamic concept, which requires assessment appropriate to each context.[35] Depending on the facts of the case, online harms may infringe upon this right.

- **Freedom of expression:** Section 16 of the Constitution states that everyone has the right to freedom of expression. Online harms discourage victims and survivors from freely using online platforms.[36] Online harms also have a secondary effect of preventing people from exercising their right to freely express themselves, without fear of retaliation, on social media platforms.

The psychological and social effects of online harms

Online harms may result in many negative psychological and social effects:

- **Psychological effects**: Online harms may affect victims and survivors psychologically in a variety of ways.[37] Online harms may cause anxiety, fear and stress.[38] Online harms may contribute to or cause depression. These psychological effects may be completely debilitating, resulting in victims and survivors losing their ability to enjoy their lives.

- **Social consequences:** Online harms may lead to the loss of enjoyment of life due to the negative effects they may have on a victim or survivor's social interactions. Online harms, due to the psychological effects that they may cause, as discussed above, may lead to victims and survivors reducing their participation in social life.[39]

The physical and offline effects of online harms

Online harms may have effects on victims and survivors' physical health and can also manifest into offline harms.

- **Effects on physical health:** The psychological effects of online harms, which have been discussed above, may result in physical health consequences. For example, prolonged stress can weaken the immune system, which makes individuals more susceptible to illness.[40] Depending on the type of online harm experienced and the social consequences of such harm, a victim or survivor may also turn to self-harming practices as a coping mechanism.[41]

---

35 General Comment No.25(2021) on Children's Rights in Relation to the Digital Environment CRC/C/GC/25 (2021) at para 12 (accessible **here**).
36 Centre for Human Rights, 'Understanding Online Gender-based Violence in Southern Africa' (2021) (accessible **here**) at page 9.
37 As above.
38 See, for example, the effect of online harassment: South African Human Rights Commission, 'Social Media Charter' (2023) (accessible **here**) at page 19.
39 M Storry and S Poppleton, 'The Impact of Online Abuse: Hearing the Victim's Voice' Victims Commissioner (2022) at pages 7 and 24 (accessible **here**).
40 F Dhabhar, 'Effects of stress on immune function: the good, the bad, and the beautiful' (2014) (accessible **here**).
41 NSPCC, 'Online abuse' (2024) (accessible **here**).

- **Offline effects:** Online harms may translate into offline harms and a further infringement of rights. For example, the non-consensual sharing of a person's private information, such as their address, may lead others who then have access to that information to stalk or physically harm the victim or survivor.[42] Further, online hate speech may be used by a person or group to incite others to engage in physical violence.[43]

- **Increased risk of exploitation:** Children may be more vulnerable to certain types of online harms, such as online predation and exploitation. Depending on the child in question, they may not fully understand the risk associated with sharing personal information or engaging with strangers.[44]

42 Storry and Poppleton above n 37 at page 16.
43 There have been various examples of this around the world. In South Africa, campaigns such as Operation Dudula have used online platforms to incite xenophobic violence (see **here**).
44 S Rutty, 'Online harms: protecting children and young people' NSPCC (2024)(accessible **here**.)

4.
An overview of the main social media platforms

Online harms occur mostly on social media platforms. Therefore, in order to fully understand online harms and how they manifest, it is necessary to understand how these platforms work. It is important to note that new social media platforms may be created and become popular at any time, and the nuances of each platform must be examined. The below section sets out the features and nuances of the platforms that are presently popular in South Africa – WhatsApp, X, Instagram, Facebook, and TikTok.

WhatsApp

WhatsApp is an application (app) through which people can send text messages, voice messages, photos, videos, documents and even locations to their contacts. Direct messages are end-to-end encrypted, meaning that only the sender and recipient can read the message.[45] In order for someone to be one of your contacts on WhatsApp, you must have the person's phone number saved, and they must also have a WhatsApp account.[46] In order to create an account, one needs to download the app and register with a valid phone number, as well as select a username. A user's contacts will be able to see their phone number and username on their profile, depending on their privacy settings.[47]

---

45 WhatsApp Help Center, 'About end-to-end encryption' (accessible **here**).
46 WhatsApp Help Center, 'How to find and invite contacts' (accessible **here**).
47 WhatsApp Help Center, 'How to change your privacy settings' (accessible **here**).

Key terminology: read receipts and blocking

Read receipts are a feature on WhatsApp that indicates to a user's contacts whether they have read their messages or not. The symbol of one grey tick next to a message indicates that the message was successfully sent, two grey ticks indicate that the message was successfully delivered, and two blue ticks indicates that the recipient has read the message.[48] However, users may change their settings so that the ticks will always be grey, regardless of whether they have read the message or not.

On WhatsApp, users have the option of **blocking** someone. When a user blocks another, the blocked user can no longer send them messages, see their status, view their profile picture, or see their contact information.[49] The blocked user does not get notified that they have been blocked.[50]

WhatsApp has several features that take it beyond an individual-to-individual messaging platform. These features must be examined to holistically understand how the app functions.

- **Disappearing messages:** Disappearing messages is an optional feature that users can turn on in their settings. Using this feature, users can set messages to disappear after 24 hours, 7 days, or 90 days.[51] After the message disappears, both the sender and receiver will no longer be able to see it. This may have adverse implications for evidence gathering in the case of an online harm, unless the victim or survivor screenshots the message before it disappeared.[52]

- **View once:** This feature allows users to send photos, videos or voice messages that disappear after the recipient opened them once. r. Most smart phones prohibit recipients from taking screen grabs of content sent through this feature. The utility of the service is limited but often used by those seeking to extort recipients as it prevents messages and images from being captured and used as evidence.

- **Groups:** Users can create groups, with up to 1024 participants, on WhatsApp.[53] In groups, the user that created the group, or participants who have been designated as admin by the creator, can adjust the group settings so that everyone can share messages or media, or otherwise so that only the admin can do so.[54] Depending on the settings of the group, group admins may have additional controls such as adding or removing members and setting group descriptions.[55] The participants in a group are able to see all of the other participants' contact information, even if they do not have that person saved as a contact on their phone. However, this can be changed depending on the privacy settings of the group.

48 WhatsApp Help Center, 'How to check read receipts' (accessible **here**).
49 WhatsApp Help Center, 'About reporting and blocking someone on WhatsApp' (accessible **here**).
50 Id.
51 WhatsApp, 'About disappearing messages' (accessible **here**).
52 A screenshot is an image of what is shown on a computer on cell phone screen. It captures what was on the screen at the moment the screenshot was taken. See Cambridge Dictionary, 'Screenshot' (accessible **here**).
53 WhatsApp Help Center, 'How to create and invite into a group' (accessible **here**).
54 WhatsApp Help Center, 'How to manage group admins' (accessible **here**).
55 Id.

- **Communities:** On WhatsApp, users can create communities, which bring related groups together under one umbrella.[56] A typical example is a suburban community, which consists of separate street groups. Every user who is in the suburb community will see what is shared in the community group, but only those in the separate street groups will see what is shared there. The community settings will dictate whether participants are able to see each other's contact information.[57] The purpose of communities is to aid in organising and managing multiple groups more efficiently.

- **Channels:** WhatsApp Channels are a new way to receive information and updates directly within the WhatsApp app from people and organisations. Channels offer a one-way broadcast model, essentially meaning that channel admins can post text, links, images, videos, and polls, but followers cannot reply or message the admin directly.[58] Users can follow any public channel of interest; unfollow at any time from the channel or info page; and forward channel updates to personal chats, groups, or WhatsApp Status.[59] Notably, other followers cannot see who is following a channel or how others interact with posts; however, channel admins may see profile names and, depending on your privacy settings, profile photos next to a reaction or poll vote.[60]

- **Statuses:** Statuses are temporary posts that disappear after 24 hours.[61] Users can post text, photos or videos as a status. The user posting the status can choose which of their contacts are able to see their status through privacy settings.[62]

If users experience abuse or inappropriate content while using WhatsApp, they can report individual contacts or groups through an integrated reporting feature in the app.[63] Once reported, WhatsApp will receive the last five messages that the reported user or group sent and then will take appropriate action based on its policies.[64]

X (previously Twitter)

X is a social media platform where users can post and interact with messages. These messages are known as tweets, although since changing its name, X refers to them as posts. Ordinary users can post short messages, up to 280 characters, including text, images, videos and links.[65] In order to create an account, users need to provide a username, phone number and email address.[66] A user's username is known as their handle.[67] Further, users may adopt a profile picture, a short description about themselves (a bio), their location, and a link to a website.[68] Anyone who visits a user's profile is able to see their profile picture, name, bio and location.

---

56 WhatsApp Help Center, 'About communities' (accessible **here**).
57 WhatsApp Help Center, 'About general chat for communities' (accessible **here**).
58 WhatsApp Help Center, 'About WhatsApp channels' (accessible **here**).
59 WhatsApp Help Center, 'How to find and follow WhatsApp Channels' (accessible **here**).
60 WhatsApp Help Center, 'About safety and privacy on channels' (accessible **here**).
61 WhatsApp Help Center, 'How to use status' (accessible **here**).
62 Id.
63 WhatsApp Help Center above n 47.
64 Id.
65 X, 'How to Post' (accessible **here**).
66 X, 'Signing up with X' (accessible **here**).
67 X, 'How to change your x Username' (accessible **here**).
68 X above n 64.

Key terminology: timelines, private messages, and public or protected tweets

On X, users each have a **timeline** which displays the tweets of those whom they follow.[69] X's algorithms also identify tweets or accounts that it deems to be popular or relevant to the user in question and adds that content to users' timelines.[70]

Users can send **private messages** or Direct messages (DMs) to those who follow them or even those who do not, depending on their privacy settings.[71] Such private messages are only visible to the sender and receiver.

Users can adjust their settings so that their **tweets are public or protected.** Tweets are by default public, meaning that anyone can see them – even people who do not have an account. However, users may also set their tweets as protected, making them only visible to their followers.[72]

X has a variety of features which can be used by users to interact with tweets. These are set out below.

- **Reposting and liking:** Users can repost tweets, which essentially means that the user re-shares an existing tweet.[73] Users can additionally add their own comment or media before reposting the tweet.[74] The user's followers will then be able to see the reposted tweet, as well as be able to go directly to the original tweet. Users can also repost their own tweets, which will bring the tweet to the user's followers' attention on their timelines again.[75] Users may also like tweets, which is an indication of admiration or endorsement. Viewers of the tweet will be able to see how many likes it has, as well as which users liked it.[76] The author of the tweet will be notified of all of the likes on their tweet.[77]

- **Replies:** A reply is a response to another user's tweet[78]. Anyone who can view the original tweet will also be able to see all of the replies made on that tweet.[79] Therefore, users will be able to see replies even if they do not follow the user who made the reply. The author of the original tweet is able to delete replies made on their tweet, and thereafter, no one will be able to see the deleted reply.[80]

69 X, 'About your for you timeline on X' (accessible **here**).
70 Id.
71 X, 'About Direct Messages' (accessible **here**).
72 X, 'About public and protected Posts' (accessible **here**).
73 X, 'How to Repost' (accessible **here**).
74 Id.
75 Id.
76 X, 'How to like a post' (accessible **here**).
77 Id.
78 X, 'About replies and mentions' (accessible **here**).
79 Id.
80 Id.

- **Mentions**: Users can mention each other in tweets or replies. The effect of mentioning someone is that it creates a link to the mentioned user's profile and notifies them about the tweet or reply in question.[81] A user who has been mentioned in a tweet can untag themselves from unwanted tweets or replies.[82]

- **Threads**: A thread on X is a series of connected tweets from one user.[83] Users can provide additional context, an update to a tweet, or an extended point by connecting multiple tweets through a thread.[84] Users will be able to see all of the tweets within a thread, as well as all of the likes, reposts, replies, and mentions from each individual tweet.[85]

On X, users are able to report tweets, replies, or private messages that are in violation of the platform's rules or terms of service.[86] For example, users can report tweets that contain abusive content. X will then review the reported content and, if it deems that the content is in violation of its policies, it will take action.[87] Action in this sense may result in anything from a warning to the perpetrating user to a suspension of their account.

It is important to note, however, that X is often unresponsive or inconsistent in addressing such reports. Users should also be encouraged to also report harmful content through independent systems such as the **Real411** and the **Media Attack Report System (MARS)**. Users should also be encouraged to save posts before reporting them by taking screenshots or screen grabs, for example, so that evidence is preserved even if the content is later deleted.

Grok

Grok is an AI assistant available to users on X. It is designed to help users complete everyday tasks such as answering questions, solving problems, and generating ideas. Interestingly, Grok is able to incorporate real-time data by searching public posts on X and, when necessary, performing live web searches. Its large language model was pre-trained on publicly available data and refined with input from human AI reviewers known as AI Tutors.[88]

When users interact with Grok, X may share their public profile information, engagement data, and interaction history with AI to enhance Grok's performance.[89] This information helps personalise responses and fine-tune Grok's language capabilities, including its tone, accuracy, and contextual understanding.[90] Users have control over these processes and can opt out of data use for model training or personalisation through their privacy settings on X.[91]

---

81 Id.
82 Id.
83 X, 'How to create a thread on X' (accessible **here**).
84 Id.
85 Id.
86 X, 'Report abusive behavior' (accessible **here**).
87 Id.
88 X Help Center, 'About Grok, Your Humorous AI Assistant on X' (accessible **here**).
89 Id.
90 Id.
91 Id.

The latest release of Grok-2 is capable of producing highly realistic and sometimes provocative images, including fabricated depictions and has done so for well-known public figures such as Donald Trump and Kamala Harris in controversial contexts.[92] X's policy of minimal content moderation, closely tied to Elon Musk's (the owner of X) broader philosophy of free expression, has drawn widespread concern as the system's ability to generate striking, misleading visuals adds a complex layer to the already difficult task of moderating harmful content online.[93] Without stronger safeguards, Grok-2 could be weaponised and used to facilitate different forms of online harassment.

Instagram

Instagram is a social media platform that enables users to share and interact with photos and videos. In order to create an account, users must be above 13 years of age, provide a valid phone number, email address, and a name (which can be real or a pseudonym).[94] Users may also elect to upload a profile picture, description of themselves, and a link to a personal or professional website.[95] Anyone who visits a user's profile is able to see their profile picture, name, bio, and website.[96]

Key terminology: followers, feeds, and public or private accounts

The concept behind Instagram is that users **follow** each other. If you follow a user, you are known as their follower.[97]

Once a user follows another, their posts will be displayed in their **feed**.[98] A user's feed is where all the posts from the people that they follow, as well as advertisements that Instagram's algorithm thinks that they would be interested in, are displayed. A user's feed is the first thing that they see when they go into the app.

Users can elect whether to make their **account private or public**. If an account is public, it means that any user who visits their profile can see what they have posted and may follow that user.[99] If a user's profile is private, it means that others will only be able to see their posts and follow them if that user has given their permission.[100]

Instagram has developed to allow users to post a variety of content and to interact with content in a variety of ways. This is set out below.

92 Times of India, 'Why Elon Musk's Grok-2 AI images is raising serious concerns' (2024)(accessible **here**).
93 Id.
94 Instagram Help Center, 'Your Profile' (accessible **here**).
95 Id.
96 Id.
97 Instagram Help Center, 'Syncing Contacts and Finding People to Follow' (accessible **here**.)
98 Instagram Help Center, 'How Instagram Feed Works' (accessible **here**).
99 Instagram Help Center, 'Differences between public and private accounts on Instagram' (accessible **here**).
100 Id.

Photos, videos, or reels: On Instagram, users can post a photo, video, or a series of photos and videos, either with or without a caption. A series of photos or videos in a singular post is known as a carousel.[101] Further, users may also record, edit and then post short videos known as reels. Users can use the reels function to merge existing video clips, record new video clips, and add music and other effects to the clips.[102] Reels may similarly be posted with or without captions.

- **Stories:** Users can share photos and videos that will disappear after 24 hours. These temporary posts are known as stories.[103] However, users can elect to add a story to their highlights, which means that the story will then be visible on the user's profile even after the 24-hour period has lapsed.[104] Users can also elect to only share a story with their close friends. Users can choose which of their followers is in their close friends list, and only those users will be able to see the story in question.[105]

- **Live:** Users can live broadcast to their followers. Everyone who follows the user will be able to see in real time whatever the broadcaster's camera is filming at that moment.[106] Once a live broadcast has ended, only the broadcaster will be able to share a replay of it or access it in their archive.[107]

- **Comments:** If a user has a public profile, anyone may comment on their posts. If the user has a private profile, only those following them may comment on their posts. Any user who is able to view the post will also be able to see all of the comments made and also be able to reply to any comment that has been made. The author of the post can delete comments and replies thereto. Users may also elect to turn off comments.[108] Thereafter, others will not be able to comment on their posts.

- **Tags and mentions:** Users can "tag"[109] other users in the photos, carousels, reels, or stories that they post.[110] If a user has a private profile, they will only be able to tag their followers. Anyone who is able to see the post will be able to see who has been tagged in said post. The user who has been tagged will get a notification of the same, and the post will be added to the tagged user's profile.[111] However, if the tagged user has a private profile, only their approved followers will be able to see posts that they have been tagged in.[112] If a user is tagged in a story, they will be able to re-post the story. Users can also mention other users in a comment. The effect of mentioning another user in a comment is that that user will then get a notification of the same, drawing their attention to the post in question.[113] If the user who posted the post has a private account, only those who follow them will be able to see the original post, even if they

---

101 Instagram Help Center, 'Sharing Photos and Videos' (accessible **here**).
102 Instagram Help Center, 'Reels' (accessible **here**).
103 Instagram Help Center, 'Stories' (accessible **here**).
104 Id.
105 Instagram Help Center, 'Create a Close Friends list on Instagram' (accessible **here**).
106 Instagram Help Center, 'Live' (accessible **here**).
107 Id.
108 Instagram Help Center, 'Turn comments on or off for Instagram posts' (accessible **here**).
109 To tag something means to use a word/s or phrase to categorise content and make it easier for users to find and engage with specific topics.
110 Instagram Help Center, 'Tagging and Mentions' (accessible **here**).
111 Id.
112 Id.
113 Id.

have been mentioned in a comment. Instagram has a range of settings to manage who can tag or mention a user in a post. For example, a user could choose to only allow tags or mentions from people that they follow.[114]

- **Direct messages:** Instagram allows users to send direct messages to other users or to a group of users.[115] These messages can include text, photos, and videos. Users can also send profiles or posts to other users.[116] Users may adjust their settings so that only users whom they follow can send them messages directly. Messages that are sent from users who do not follow will be added to a requests folder – the user will have to accept the message before it appears in the user's chats.[117]

- **Explore:** Besides seeing posts, stories, and live broadcasts from users that one follows, Instagram also offers a feature known as the explore page. The explore page is curated by Instagram's algorithms, pulling together posts from users that the user does not follow, based on users that they do follow and posts that they interact with.[118] For example, if a user primarily follows users who post travel content, their explore page will most likely comprise similar travel content from other users. The explore page is also where users can search for profiles or specific types of content.[119]

If users think that a post, message, or story does not follow Instagram's community guidelines or terms of use, they can report it using the app's embedded reporting features.[120] If Instagram finds that the content violates its community guidelines or terms of use, it will hide or remove that content from the app.[121] Should Instagram find that there was no violation, the user will have an opportunity to request another review.[122] Users can also report a profile if that profile posts content that it should not, is impersonating someone else, or is a child below the age of 13.

Facebook

Facebook is a social media platform that allows users to share updates, photos, and videos with their friends and join communities of interest. Instagram and Facebook are owned by the same company, Meta, and as a result share many of the same features. In order to create an account, a user must provide their first and last name, phone or email address for verification, a password, their gender, and their date of birth.[123] Users must be at least 13 years old.[124] Users may also provide a profile picture, a short description of themselves, and a cover photo.[125] Visitors to a user's profile will be able to see their name, cover and profile pictures, and description. However, depending on the user's privacy settings, visitors may also be able to see additional information such as their contact details and birth date.[126]

---

114 Instagram Help Center, 'Who can mention or tag you on Instagram' (accessible **here**).
115 Instagram Help Center, 'Send, View and Manage Messages' (accessible **here**).
116 Id.
117 Id.
118 Instagram Help Center, 'Exploring Photos and Videos' (accessible **here**).
119 Id.
120 Instagram Help Center, 'Report a post or profile on Instagram' (accessible **here**).
121 Instagram Help Center, 'What to do if Instagram didn't remove content you reported' (accessible **here**).
122 Id.
123 Facebook Help Center, 'Create a Facebook Account' (accessible **here**).
124 Id.
125 Id.
126 Facebook Help Center, 'Your Privacy' (accessible **here**).

Key terminology: friends, feeds, and timelines

On Facebook, a user's connections are known as their **friends**. Friending someone refers to adding someone as a connection on Facebook. In order for users to become friends on Facebook, one of them must add the other, and that user must accept the friend request. Once added as a friend, users will be able to see each other's activity on Facebook.[127]

A user's **feed** is where all of the content that their friends, and the pages and groups that they follow, post will appear.[128] A user's feed is the landing page of Facebook, and is the first thing that they will see when opening the app.

A user's **timeline** is a section of their profile where one can see all of their posts and activities in chronological order. Posts from users' friends that they are tagged in will also be displayed on their timeline.

Facebook's main features work very similarly to those of Instagram. Users can post photos,[129] videos, reels,[130]stories,[131]and live broadcast.[132] Other users can then share, comment on, or like this content. Differently to Instagram, Facebook users may also post text without an accompanying photo or video. The visibility of Facebook posts is dependent on the author's privacy settings – users may dictate that only their friends can see their posts, or otherwise that their posts are visible to anyone who views their profile. While many of the features of Facebook are similar to Instagram, there are some different features that are set out below.

- **Tagging:** Similarly to Instagram, users may tag others in their posts or in comments.[133]However, when users tag someone on Facebook, that post will then be visible to both the user tagged and their friends.[134]. Users can opt to hide the post in question from the friends of the user tagged.[135] Users can adjust their settings so that they can review tags before they are allowed to appear on their timeline, feed, or elsewhere on Facebook.[136]

127 Facebook Help Center, 'Friending' (accessible **here**).
128 Facebook Help Center, 'How Feed Works' (accessible **here**).
129 Facebook Help Center, 'Photos' (accessible **here**).
130 Facebook Help Center, 'Reels' (accessible **here**).
131 Facebook Help Center "Stories". (Accessible **here**.)
132 Facebook Help Center, 'Go live on Facebook' (accessible **here**).
133 Facebook Help Center, 'Tagging' (accessible **here**).
134 Id.
135 Id.
136 Id.

- **Facebook Dating:** Facebook Dating is essentially a dating app under the Facebook umbrella. At the moment, Facebook dating is not available in South Africa.[137] Through users' Facebook profiles, they can create a Dating profile.[138] Users with a Dating profile may then like another user who also has a Dating profile, and, once that user has liked them back, they can direct message each other through their Dating profiles.[139] While a user's Dating profile is created through their Facebook account, it is kept separate – a user's Facebook friends and anyone outside of Facebook Dating cannot see the user's Dating profile.[140] Users are able to delete their Dating profiles without deleting their Facebook account, but deleting their Facebook account would delete their Dating profile.[141]

- **Events:** All users can create online or in-person events on Facebook.[142] Users can specify the relevant details of the event, such as the date, time, and location. The event can either be set as private or public. If it is private, only those whom the event creator invites will be able to see the event itself as well as the information about the event. If it is public, all users will be able to see it and the information about the event.[143] Users are also able to search for events that are happening near them, specific types of events, and events that their friends are going to if those events are set as public.[144]

- **Pages:** A Facebook Page is a public profile that is created for businesses, organisations, public figures, brands, or anyone seeking to promote themselves through Facebook.[145] Users can like or follow a page to receive updates from that page, which will appear on their feed. Pages can post or create events in the same way that individual users can, which will then be visible to those who follow the page.[146] Importantly, Pages are public by default; therefore, anyone on Facebook can view content on a Facebook Page without needing permission.

- **Groups:** On Facebook, users can create groups centred on different themes or interests.[147] The user who created the group is known as the Group Admin, who then controls the functioning of the group.[148] The Group Admin may also delegate this control to other members of the group. Depending on the group's settings, it may be an open or closed group. If it is an open group, any user may join. If it is a closed group, the Group Admin will need to approve all requests to join the group.[149] Once admitted to the group, members can see posts and events created by the group, which will appear on their feeds. A group may post or create events in the same way as an individual user, and all those who are part of the group will be able to see its posts on their feeds.

---

137 Facebook Help Center, 'Facebook Dating' (accessible **here**).
138 Id.
139 Id.
140 Id.
141 Id.
142 Facebook Help Center, 'Events' (accessible **here**).
143 Facebook Help Center, 'Control who sees or joins your Facebook event' (accessible **here**).
144 Facebook Help Center, 'Find Facebook events you might be interested in' (accessible **here**).
145 Facebook Help Center, 'Pages' (accessible **here**).
146 Id.
147 Facebook Help Center, 'Groups' (accessible **here**).
148 Facebook Help Center, 'Join a Facebook group' (accessible **here**).
149 Id.

- **Marketplace**: Facebook Marketplace is a platform embedded within Facebook where users can buy and sell items.[150] Users interested in buying items on Marketplace can search by category or for something specific.[151] Users can filter the results by price range or by location. In selecting an item, a user will be able to directly message the seller as well as view the seller's profile.[152]If a user wishes to sell something on Marketplace, they can list the item.[153] They will be required to give a description of the item, the price it is being sold for, and an approximate location of the item.[154] All listings are public and can be seen by anyone on Marketplace.[155]

On Facebook, users may make use of the app's embedded reporting function to report abusive content.[156] Once content is reported, Facebook reviews it and removes it if it goes against its community standards.[157]

Meta AI

Meta AI is Meta's generative AI assistant and has been integrated across its major platforms, including WhatsApp, Instagram, and Facebook. The AI assistant helps users ask questions, generate content, brainstorm ideas, or access quick recommendations.[158] Users can chat directly with Meta AI in one-on-one or group conversations, receive personalised responses, and even create AI-generated images.[159] These interactions are facilitated via natural language and, in supported regions, voice input. Similar to Grok, Meta AI responses are generated based on user prompts, with Meta collecting only those messages explicitly shared with or addressed to @Meta AI.[160] Personal messages and calls on WhatsApp remain protected with end-to-end encryption.[161]

Meta AI's design, especially its data collection, personalisation methods, and default sharing mechanisms, has raised significant privacy and safety concerns. While Meta states that only shared prompts are used for AI responses, some implementations have allowed public visibility into user queries and outputs, leading to unintended data exposure.[162] Another area of concern is Meta AI's image generation features, which have led to questionable content outputs similar to those noted in Grok.[163]

150 Facebook Help Center, 'Marketplace' (accessible **here**).
151 Facebook Help Center, 'Find items to buy on Facebook Marketplace' (accessible **here**).
152 Id.
153 Facebook Help Center, 'Selling on Marketplace' (accessible **here**).
154 Id.
155 Id.
156 Facebook Help Center, 'Report Content on Facebook' (accessible **here**).
157 Facebook Help Center, 'Reporting Abuse' (accessible **here**).
158 WhatsApp Help Center, 'About Meta AI' (accessible **here**).
159 Id.
160 Id.
161 Id.
162 A Silberling, 'The Meta AI app is a privacy disaster' (2025) TechCrunch (accessible **here**.)
163 Id.

TikTok

TikTok is a social media app that enables users to create and share short-form videos. In order to create an account, users need to provide a username, and valid email address and phone number.[164] and must be at least 13 years old.[165] Users may additionally select a profile picture and nickname. Visitors to users' profiles will be able to see their username, profile picture, and nickname.[166] Users are able to choose between a private and a public account. If a user chooses a private account, they will have control over who is able to follow them, watch their content, and who can view who the user follows.[167] On the other hand, a public account will be visible to anyone on or off TikTok, unless the user adjusts their privacy settings.[168]

Key terminology: hearts and for you

Hearts are how users show their admiration for a post. It is the equivalent of a like on Facebook. Users will get notified when another user hearts their content, and other users are able to see who hearted a post.[169]

Users **for your** pages are personalised feeds of content, based on the user's interests and engagement. It is the first thing that users see when they open the app, and is comparable to the explore page on Instagram.[170]

Users can elect to either create videos through the app or otherwise upload existing videos from their devices.[171] Users may then post the video or keep it in drafts to post at a later date.[172] Users may elect to add a caption to the video or not.[173] Once posted, the visibility of the video is dependent on whether the user has elected to have a public or private profile, and their privacy settings, as set out above. As with other social media platforms, there are particular features of the app that require explanation.

- **Tagging and comments:** Users can personalise their privacy settings to specify who can comment on their videos.[174] For example, users may elect to only allow users whom they follow and who follow them back to comment on their videos. Comments are then visible to all users who are able to view the original post. Users may also elect not to allow comments for specific videos.[175] Users can similarly control who may tag them in posts, stories, and comments.[176] Once tagged, all those who can view the original post, story, or comment will be directed to the tagged user's profile.[177] Whether they can then view the content of the tagged user's profile is dependent on the tagged user's privacy settings.[178]

---

164 TikTok, 'Creating an account' (accessible **here**).
165 TikTok, 'Youth Safety and Well-Being' (accessible **here**).
166 TikTok, 'Choosing between a private or public account' (accessible **here**).
167 Id.
168 Id.
169 TikTok, 'TikTok Glossary Part 1' (accessible **here**).
170 TikTok, 'For You' (accessible **here**).
171 TikTok, 'Making a post' (accessible **here**).
172 Id.
173 Id.
174 TikTok, 'Comment' (accessible **here**).
175 Id.
176 TikTok, 'Mentions on TikTok' (accessible **here**).
177 Id.
178 Id.

- **TikTok Stories:** TikTok Stories allow users to share photos and videos with their followers for 24 hours (or to anyone, if the user's profile is public).[179] Users who are able to view these temporary posts are then able to react and send replies to the story, which only the author of the story will be able to view.[180]

- **Direct messages:** Direct messaging allows users to privately share text, photos, or videos with each other.[181] Users can send each other public posts of other users, as well as novel text, images, or videos. Users can control who can directly message them through their privacy settings.[182]

- **Duets:** Users with public accounts can Duet with other users who also have public accounts.[183] A Duet contains two videos that display in a split screen, which play at the same time.[184] Users can control in their privacy settings who is permitted to Duet their videos.[185] For example, users can choose to only allow users whom they follow and who follow them back to Duet with their videos. If a user Duets with another's video, the Duet will only appear on the profile of the person who recorded the Duet.[186]

- **Stitch:** Stitch allows users to combine existing videos on TikTok from public users with one that they create.[187] A Stitch is different from a Duet in that it displays in a sequence. First, the original video will play, followed by the stitched content. In a Duet, the two videos display side by side. Similarly to Duets, users can control through their privacy settings who can Stich their videos.[188] The Stitched video will only appear on the profile of the user who recorded the Stitch.[189]

TikTok users may report content through the app's embedded reporting mechanism.[190] If the content violates TikTok's policies, they will remove the content after they have reviewed it.[191] Depending on the legality of the content, TikTok may elect to remove it globally or only in the country where it is illegal.[192]

Community standards, reporting mechanisms, transparency of processes, and appeal mechanisms all differ significantly from platform to platform. Each platform has its own set of guidelines and procedures for handling potentially harmful content, which can lead to a lack of consistency in how issues such as hate speech and incitement are addressed. For example, some platforms may have more stringent policies and quicker response times, while others might be less transparent about their processes and slower to act on user reports.

---

179 TikTok, 'TikTok Stories' (accessible **here**).
180 Id.
181 TikTok, 'Direct Messages' (accessible **here**).
182 Id.
183 TikTok, 'Duets' (accessible **here**).
184 Id.
185 Id.
186 Id.
187 TikTok, 'Stitch' (accessible **here**).
188 Id.
189 Id.
190 TikTok, 'Reporting Illegal Content' (accessible **here**).
191 Id.
192 Id.

The Real 411

Some civil society organisations are developing tools to ensure consistency in the way that online harms are reported and thereafter processed. For example, users may report problematic content to Real411. Real411 is an independent platform where South African users can lodge complaints about online harms.[193] This platform allows users to report content they believe violates community standards or legal provisions related to hate speech and incitement, providing an additional layer of oversight and accountability beyond the internal mechanisms of individual platforms.

193 Real 411 (accessible **here**).

5.
Existing legislation

This section outlines the relevant legislation that addresses various aspects of online harms and will assist SAPS in navigating different cybercrimes. Specifically, this section will outline and give context to the following legislation:

- The Electronic Communications and Transactions Act;
- The Cybercrimes Act;
- The Protection from Harassment Act;
- Prevention and Combating of Hate Crimes and Hate Speech Act; and
- The Criminal Law (Sexual Offences and Related Matters) Amendment Act

Electronic Communications and Transactions Act

The Electronic Communications and Transactions Act 25 of 2002 (ECTA)**194** covers email, internet, SMS, and other electronic messages (but not regular voice calls between two people). Chapter XIII introduces the criminal offences:

- **Section 86** covers hacking, password trading, tapping into data flows, and spreading viruses or launching denial-of-service attacks.
- **Section 87** criminalises the use of a computer to threaten someone for money or commit fraud (e.g. blackmailing someone to stop a cyberattack).
- **Section 88** imposes criminal offences on anyone who helps or encourages someone to commit the abovementioned crimes.
- **Section 89** states that offenders can face fines or imprisonment. Offenders can face imprisonment for up to 12 months for minor offences and 5 years for serious ones.

194 Electronic Communications and Transactions Act 25 of 2002 (accessible **here**).

The Cybercrimes Act

The Cybercrimes Act 19 of 2020[195] extends the offences outlined in ECTA and includes a broader range of activities related to the illegal use of computer systems.

- Chapter 2 details:
  - various forms of cybercrimes;
  - malicious communications;
  - sentencing; and
  - protective orders for victims of harmful communications.
- Chapter 4 outlines the powers of law enforcement to investigate, search, access, or seize relevant data.
- Chapter 8 sets out the obligations of electronic communications service providers and financial institutions to report cybercrime offences and preserve relevant information for investigations.

The Cybercrimes Act criminalises various activities:

- **Section 2** covers unlawful access to computer systems or data storage mediums.
- **Section 3** deals with the unlawful interception of data.
- **Section 4** addresses the unlawful use of software or hardware tools.
- **Section 5** criminalises the unlawful interference with data or computer programmes.
- **Section 6** focuses on the unlawful interference with computer data storage mediums or systems.
- **Section 7** criminalises the unlawful acquisition, possession, provision, receipt, or use of passwords, access codes, or similar data.
- Other sections address cyber fraud, cyber forgery and uttering, cyber extortion, and theft of incorporeal property.

195 Cybercrimes Act 19 of 2020 (accessible **here**).

Some popular cybercrimes to look out for

South Africa has been ranked as the number one target for cyberattacks in South Africa. It is therefore important to be aware of some of the ways in which crimes have been committed online. Here are some examples of popular cybercrimes in the country:

- Ransomware: A type of malicious software that locks your files or systems and demands money (a ransom) to unlock them.
- Romance scams: Scammers pretend to be in love with someone online to trick them into sending money or personal details.
- Sextortion Criminals threaten to release private or explicit photos unless the affected person pays money or sends more content.
- Hacking: Unauthorised access to computers or networks to steal, change, or destroy data.
- Phishing scams: Fake emails or messages that trick people into giving away personal info like passwords or banking details.

The Protection from Harassment Act

The Protection from Harassment Act 17 of 2011[196] addresses various forms of harassment, including online harassment or cyberbullying. It was introduced to provide victims and survivors of harassment with a legal tool to protect themselves and to seek relief, particularly through obtaining protection orders. The Act defines harassment broadly, including both physical and electronic forms of abuse. The law also covers behaviours that cause harm, either directly or indirectly, through communication or by sending messages that induce fear, harm, or distress.

While the Act offers a framework to deal with online harassment, it has limitations:

- **Group harassment**: The Act is individualistic, meaning it does not cater to instances where harassment is perpetrated by a group of people acting together on platforms like social media.

- **Organisational harassment**: It does not protect organisations or groups targeted by harassment. Each individual would have to apply for a protection order, which can be costly and time-consuming.

196 Protection from Harassment Act 17 of 2011 (accessible **here**).

What to do when someone seeks a protection order

- Help the complainant complete an affidavit describing the abuse or threats.
- Provide the application form for a protection order.
- Make sure that they read the document titled "Notice to Complainant in a Case of Domestic Violence" before filling out the form. This document explains their rights, how to protect themselves, and what steps they can take.
- Encourage the complainant to include supporting affidavits from witnesses or others with knowledge of the abuse.
- Ensure the complainant knows where the nearest Magistrate's Court is located.
- Advise them to hand in all documents (application form + affidavits) to the court clerk.
- Advise the complainant that the court will consider the application immediately and may issue an interim protection order.

Tips:

- Keep printed copies of this notice available at the station or direct them to the official **download link** if online access is available.
- Offer to help witnesses draft their affidavits clearly.
- If the complainant is vulnerable or afraid, offer to escort them to court or arrange safe transport.

Prevention and Combating of Hate Crimes and Hate Speech Act

The Prevention and Combating of Hate Speech Act 16 of 2023[197] criminalises hate speech and incitement, including that which occurs online.

- **Section 4(b)** identifies hate speech by way of electronic communication is specifically identified as a form of prohibited conduct. Electronic communication is widely defined to mean a communication by means of data messages. Therefore, hate speech and incitement that is communicated through social media platforms is criminalised.

Important to note: While the Act has been signed by the President, it is not yet in force.[198]

197 Prevention and Combating of Hate Crimes and Hate Speech Act 16 of 2023 (accessible **here**).
198 See updates on the enactment of the Act (**here**).

The Criminal Law (Sexual Offences and Related Matters) Amendment Act

The Criminal Law (Sexual and Related Matters) Amendment Act[199] was created to strengthen South Africa's laws on sexual offences. Its main purpose is to protect all individuals, especially groups like women, children, and persons with disabilities from sexual abuse, exploitation, and related harms.

This Act is especially relevant when dealing with sexual offences committed on digital platforms. These include:

- **Section 19A** which criminalises the electronic distribution of private sexual images without consent.
- **Sections 20(3)–(4)** which cover electronic grooming, where offenders use digital communication to lure children into sexual activity.
- **Section 11A** which covers the exposure of children to pornography, including via electronic means like messaging apps or websites.

Tips on handling sensitive matters relating to sexual offences

- **Avoid interrogation-style questioning:** Victims and survivors may already feel blamed or judged. Use open-ended, non-leading questions.
- **Validate their experience.**
- **Be mindful of body language and tone:** Remain calm and respectful.
- **Recognise stigma and silence:** In many communities, survivors fear social rejection or retaliation. Be discreet and protective.
- **Avoid assumptions based on gender, age, or relationship status:** GBV affects people across all demographics, including men and LGBTQ+ individuals.
- **Be aware of traditional norms:** In some cases, survivors may be pressured to reconcile with perpetrators. Respect their autonomy.
- **Secure devices respectfully:** If a phone or laptop is needed for evidence, explain why and ensure it is returned promptly.
- **Protect privacy:** Never share or discuss sensitive digital content with unauthorised personnel.
- **Make sure the case is logged as a sexual offence** and not downgraded to assault or domestic dispute.
- **Connect survivors to support services:** NGOs like Rape Crisis, Lifeline, and Thuthuzela Care Centres offer counselling and legal aid.

199 Criminal Law (Sexual and Related Matters) Amendment Act 13 of 2021 (accessible **here**).

6.
Emerging trends and case studies in South Africa

Currently, South Africa is witnessing several trends in online harms. Here are some cybercrime trends to consider:

- A major trend is the surge in phishing and business email compromise attacks. These scams now often involve AI-generated messages and deepfake content, making them more convincing and harder to detect. South African organisations and individuals are frequently targeted through email spoofing, WhatsApp scams, and voice/video phishing, leading to identity theft and financial loss.

- Deepfake technology and AI-driven harassment are also serious threats. Perpetrators use AI to create realistic fake videos or impersonate individuals, often for blackmail or to manipulate public opinion. These cases are difficult to investigate without advanced forensic tools because it is becoming harder to differentiate between what is real and what isn't.

- Doxxing has become another widespread issue. In 2025, a South African user known as "Twatterbaas" was doxed by a journalist, revealing personal details and sparking public backlash.[200] In a country with high rates of violence, such exposure can be life-threatening.

- There is a deeply troubling rise in CSAM and online sexual exploitation. In one shocking case, a couple in Midrand was found in possession of over 10 million images and videos of CSAM, along with large sums of cash believed to be proceeds from the crime.[201] Research has also found that offenders now often pose as fellow gamers or allies, using flattery, gifts, and emotional manipulation to build trust.[202] Once a relationship is established, they may introduce inappropriate content, solicit personal information, or even arrange real-world meetings.

---

200 P Maritz, 'Doxxed for dissent' (2 May 2025)(accessible **here**)
201 S Masilela, 'Update: Midrand man arrested for distributing child pornography after police find 10 million images and videos' IOL (17 January 2025)(accessible **here**).
202 Daybreak, 'The Alarming Rise of CSAM and Child Exploitation in South Africa' (2024) (accessible **here**).

7.
Practical considerations

In order to adequately deal with complaints regarding online harms, it is important to take practical considerations into account. First, a police official can seek to investigate whether there are any patterns of harm to be aware of as opposed to a single incident or event. Second, each complaint must be assessed within the context within which it occurs. These practical considerations further extend to the unique challenges and opportunities present when dealing with online harms as a result of their occurrence in the online world.

Recognising patterns of online harm

Sometimes, a single incident of online harm may not seem serious enough to qualify as a criminal offence. However, when looking at the bigger picture, repeated incidents against the same person may show a clear pattern of unlawful behaviour.

For example, while a single offensive message on its own, it may not meet the legal threshold for harassment, reading a series of messages, or other harmful actions online, can reveal an organised or sustained campaign of harassment against victims and survivors.

Practical tool for recognising patterns of online harm

As members of SAPS, it then becomes important to:

- **Look beyond single incidents:** Consider whether the harm is part of an ongoing pattern.
- **Recognise group behaviour:** Sometimes the harm is carried out by several people acting together which can make the impact more damaging.
- **Watch for ripple effects:** One harmful post may be designed to incite others to join in, spreading abuse even further. Even if the original post seems minor, it can lead to significant harm when it triggers wider attacks.

Identifying these patterns and networked behaviours will enable members of SAPS to can gather stronger evidence of unlawful conduct, link incidents that might otherwise appear unrelated, and protect victims and survivors from ongoing or escalating harm.

Context is key

When investigating online harms such as harassment, intimidation, hate speech, or cyberbullying, it is important to consider the full context of the people involved, not just the words or images in isolation.[203] The language used, cultural references made, and the online presence of both the perpetrator and the victim or survivor can all affect how harmful the conduct is to the victim or survivor, what the perpetrator intended, and how the act is likely to be interpreted by the courts under South African law.

This approach aligns with legislation including the Cybercrimes Act[204] and the Protection from Harassment Act.[205]

Words, slang, or jokes may carry different meanings across South Africa's languages and cultures. A post that appears harmless to someone unfamiliar with the context may, in isiZulu, Sesotho, Afrikaans, or township slang, carry clear threats or cause deep humiliation. Members of SAPS should assess the communication holistically, considering the relationship between the parties, any prior complaints, and whether the language is coded or used to rally others.

The visibility of the harmful post or message is also important. The wider the reach, the greater its potential impact on the victim or survivor.[206] Members of SAPS should check how many followers or friends both the perpetrator and complainant have, and how widely the post was shared, liked, or commented on. This is especially relevant on widely used platforms such as WhatsApp, Facebook, TikTok, and X. Public posts or group messages may escalate harassment, but even private direct messages can indicate serious threats if abusive or intimidating.

---

203 P Sekati, 'Context matters: The importance of local nuance in online moderation' (2023) ALT Advisory. (accessible **here**).
204 Cybercrimes Act above at n 193.
205 Protection from Harassment Act above at n 194.
206 World Economic Forum above at n 1.

Looking at the full context, including online behaviour patterns, cultural background, and prior history, also helps to establish whether the conduct was deliberate, reckless, or accidental. Members of SAPS should take note if the perpetrator is inciting others to join in, as this can turn individual harassment into coordinated online mobbing.

Certain phrases may seem vague but can signal intimidation or threats when viewed in local context. For example, expressions such as "uzobona" ("you will see"), "ons kry jou nog" ("we'll still get you"), or "your time is coming" may indicate ongoing intimidation. Using animal names such as "dog," "pig," or "cockroach" is often meant to dehumanise.

Case study: *Kunene and Another v Malema*[207]

In August 2025, the Gauteng High Court affirmed the Equality Court's finding that calling Julius Malema a "cockroach" in a televised interview constituted hate speech under section 10 of the Promotion of Equality and Prevention of Unfair Discrimination Act. Although it overturned the earlier ruling that terms like "little frog" and "criminal" were hateful, the court held that "cockroach," having strong genocidal dehumanisation connotations, crossed the line. The court ordered Kenny Kunene to issue a written and oral public apology, interdicting him from using the term again.

This case affirms that certain words, when understood in their South African historical and cultural context, can amount to hate speech or unlawful expression.

Some useful ways to include context within an investigation would be to document how visible harmful content is by capturing screenshots, follower counts, post reach, and WhatsApp group sizes. Recording cultural or linguistic factors and, if necessary, consulting a colleague or interpreter to clarify slang or coded threats.

A further example of a harmful cultural phrase is the term "necklacing". In South Africa, necklacing refers to placing a petrol-soaked tyre around someone's neck and setting it on fire. This practice was pronounced during the mid-1980s during the anti-Apartheid resistance.[208] For South Africans, the use of the phrase invokes feelings of fear and violence. In contrast, those who do not know the cultural context of the term may take it to refer to jewellery. They would therefore not be impacted by the term in the same way. Therefore, the use of the term in the South African context may clearly show violent intention on the part of the perpetrator.

207 Kunene and Another v Malema [2025] ZAGPJHC 742 (accessible **here**).
208 R Moosage, 'A prose of ambivalence: liberation struggle discourse on necklacing' (2010) 36(1) Kronos 136. (accessible **here**).

Challenges in addressing online harms

Addressing online harms presents unique challenges for police officers. Unlike traditional crimes, digital offences can be rapid, anonymous, and far-reaching, often spreading across multiple platforms and borders before law enforcement can respond. Investigating these harms requires specialised skills, timely evidence preservation, and careful assessment of the context, intent, and impact on victims and survivors.

Collecting and preserving digital evidence

Investigating online harms requires timely and accurate evidence collection.[209] Unlike physical crimes, digital content can be quickly deleted, altered, or shared widely, making preservation critical. According to the United Nations Office on Drugs and Crime "digital evidence is volatile and fragile and the improper handling of this evidence can alter it", therefore, officers must capture screenshots, screen recordings, metadata, and platform-specific details before the content disappears.[210] Delays in preserving evidence can significantly weaken a case, as online posts may "go viral" or be removed before investigators can act.

Investigating encrypted communications

Many perpetrators use encrypted messaging apps or virtual private networks (VPNs) to conceal their identity.[211] While the Cybercrimes Act provides mechanisms for obtaining user data, officers may encounter technical limitations or lack access to digital forensics tools at local stations.[212] Decrypting communications or linking anonymous accounts to real individuals often requires specialised units, adding complexity to investigations.

Jurisdictional issues in cross-border harms

Online harms often cross national borders, as platforms hosting abusive content may be located outside South Africa. Officers may need to rely on mutual legal assistance treaties or international cooperation to obtain data or take action against offenders abroad. These processes can be slow and cumbersome, leaving victims and survivors without immediate protection. The global nature of online platforms also complicates enforcement of local laws and removal of harmful content.[213]

---

209 United Nations Office on Drugs and Crime, 'Handling of digital evidence' (accessible **here**).
210 Id.
211 EM30 Tech, 'The Dark Side of Messaging Apps: How Criminal Networks Exploit Telegram' (23 October 2024). (accessible **here**).
212 Cybercrimes Act above n 193 at section 45.
213 M Melck 'Cybercrime and Jurisdiction – where do we stand?' DML (25 June 2018) (accessible **here**.)

Procedural delays

Procedural delays can occur at multiple stages, from reporting to escalation to specialised units. Victim and survivors may perceive a lack of urgency, particularly when harm seems intangible compared to physical crime. Police officers must often balance frontline response with proper legal procedure, evidence preservation, and escalation protocols. Delays can also be exacerbated by limited cybercrime training or staffing constraints at local police stations, which can impact the speed and effectiveness of investigations.

Practical consideration: Assisting with protection orders

As outlined above, when assisting with protection orders it is important to provide clear guidance, ensure that complainants understand their rights, and encourage complainants to collect and as much evidence as possible online and through supporting affidavits.[214]

Cybersecurity and prevention strategies

Public awareness campaigns

Educating the public is a cornerstone of preventing online harms. SAPS has collaborated with various stakeholders to promote online safety. For instance, SAPS has provided guidelines on cybercrime prevention, advising the public to use strong passwords, secure devices with firewalls and antivirus software, and be cautious on social media platforms.[215]

Additionally, the Internet Safety Campaign Africa (ISC Africa) has been instrumental in raising awareness about online safety across the continent.[216] Their campaigns focus on innovation, collaboration, and transformation to ensure a safe online environment for all. However, a study published in 2025 highlighted that while there is an increasing awareness of cybercrime, many South Africans still lack basic online safety practices.[217] Professor Modise highlighted that "studies suggest South Africa has a high prevalence of cybercrime, with reports ranking it among the worst affected countries globally".[218] This underscores the need for continuous and targeted public education efforts to curb the effects of cybercrime on communities most at risk.

214 See above at pages 27-28.
215 South African Police Service, 'Top 10 cybercrime prevention tips' (accessible **here**).
216 Read more about the Internet Safety Campaign (**here**).
217 J Modise, 'The Public's Awareness of Cyber Crime and Online Safety Practices In South Africa' (2025) 2 MRS Journal of Multidisciplinary Research and Studies 14 (accessible **here**).
218 Id.

Collaboration with tech platforms

Effective prevention of online harms requires collaboration between law enforcement and technology platforms. In a notable initiative, SAPS partnered with private sector entities to combat tech-savvy criminals utilising smartphones and AI.[219] This partnership aims to outwit criminals by leveraging the expertise and resources of both public and private sectors.

Such collaborations are essential, as they enable law enforcement to access tools and data that can aid in the detection and prevention of online harms. Moreover, working with tech platforms ensures that there is a coordinated response to emerging threats thereby enhancing the overall effectiveness of cybersecurity efforts and cybercrime prevention.

Staying informed

As rapid developments take place in the digital realm, it becomes exceedingly important that those in law enforcement remain updated on the trends and nuances associated with online harms. While it may not be expected to have a comprehensive understanding of the technicalities of AI, for example, staying informed not only gives understanding to the harmful extents of online harms, but can also enable members of SAPS to be more responsive and better positioned to address crimes resulting from online harms as they arise.

Practical resources and training guidelines

- **HJN Training Academy:** Provides accredited safety, security, and ICT programmes that empower students to fight crime, close tech skills gaps, and build stronger communities across Southern Africa.
  - Contact: headoffice@hjntraining.co.za / 078 177 4023 / 0860 995 220
- **South African Human Rights Commission (SAHRC):** Includes advisory documents and resources such as the **Social Media Charter** on addressing online harms.
  - Contact: **info@sahrc.org.za** / 011 877 3600
- **ALT Advisory:** Publishes analysis, toolkits, and policy guidelines on digital rights, online safety, and legislative developments.
  - Contact: **info@altadvisory.africa** / 010 822 7860
- **SLAPPBack.org.za:** Provides resources and guidelines to help navigate common forms of weaponisation of court processes in South Africa.
  - Contact: **info@mma.org.za** / 011 788 1278
- **Tech4Law:** Legal technology newsletter covering cybersecurity, digital transformation, and online harms in law practice.
  - Contact: **editor@tech4law.co.za**
- **Lexinfo CC Attorney Alert:** Provides legal updates across South African law, including digital rights and online harms.
  - Contact: **alert@lexinfo.co.za**

219 Cape Chamber of Commerce and Industry, 'Groundbreaking public-private partnership between SAPS and private sector to outwit tech-savvy criminals' (7 August 2025)(accessible **here**).

8.
Conclusion

The rapid growth of digital technologies has transformed how people interact, but it has also created new avenues for harm, exploitation, and abuse. Online harms are not isolated to cyberspace, they have profound real-world consequences that affect personal dignity, safety, and equality. For members of SAPS, addressing these harms requires both a firm grasp of the legal framework and a trauma-informed, context-sensitive approach.

This handbook has highlighted the varied forms of online harm, ranging from cyberbullying and cyberstalking to hate speech, non-consensual sharing of intimate images, child sexual abuse material, financial cybercrime, and data breaches, and the ways in which they infringe constitutional rights. It has underscored how these harms disproportionately impact certain groups and how intersectionality shapes victims and survivors' experiences.

Context matters in determining intent, impact, and accountability in online harm cases. Likewise, the evolving challenges faced by police, such as preserving digital evidence, investigating encrypted communications, and responding to cross-border offences, demand enhanced technical capacity and close cooperation with technology platforms.

Effective policing in the digital space also calls for prevention, not just reaction. Public awareness campaigns, proactive monitoring, and partnerships with social media companies are critical in reducing harm before it occurs. Sharing the handbook with colleagues, arranging workshops with experts and sharing the tips and guidelines in Annexure A. Further, these strategies can protect victims and survivors, hold perpetrators accountable, and build public trust in law enforcement's ability to safeguard both physical and online communities.

Ultimately, the same principles that guide policing in the physical world, respect for human rights, fairness, diligence, and compassion, must guide responses to harms in the digital world. The insights, tools, and strategies set out in this handbook aim to bridge the gap between the physical and digital world, assist SAPS in confronting these modern threats, and ensure that the promise of a free and open internet does not come at the cost of safety and justice.

Annexure A: Handling complainants for online harm: Key steps

(L) – Listen and Acknowledge	**(A)** – Advise on Immediate Steps
(**A)** – Assess Immediate Risk	**(P)** – Protect the Complainant's Privacy
(**R)** – Record the Complaint Thoroughly	**(P)** – Provide Follow-Up and Support
(C) – Clarify the Process	

Step 1: Listen and Acknowledge

- Let the complainant speak without interruption.
- Validate their distress and affirm their right to safety and dignity.

Step 2: Assess Immediate Risk

- Is the complainant in physical danger?
- Are there threats that need immediate intervention (e.g., suicide ideation, threats of violence)?
- Escalate if necessary.

Step 3: Record the Complaint Thoroughly

- Gather all relevant details such as the:
    - Nature of the harm
    - Timeline of events
    - Known identity of the perpetrator (if any)
    - Platforms involved
- Advise the complainant to gather copies/screenshots of communications as soon as possible to prevent the risk of content being deleted
- Ask for as much information as possible including copies/screenshots of communications.

Step 4: Clarify the Process

- Explain:
    - What actions can be taken (investigation, warnings, arrests)
    - The potential outcomes
- The complainant's rights

Step 5: Advise on Immediate Steps

- How to block/report the offender.
- How to secure accounts (password changes, 2 Factor Authentication).
- How to document ongoing incidents.

Step 6: Protect the Complainant's Privacy

- Avoid disclosing sensitive information to third parties.
- Minimise unnecessary exposure of content provided by the complainant especially in cases of image-based abuse.

Step 7: Provide Follow-Up and Support

- Keep the complainant informed of progress
- Refer them to support groups, counselling, or legal aid if needed
- Encourage reporting further incidents

Annexure B: Example of a poster that can be placed within SAPS offices

Annexure C: Internet user guide for beginners

Topic	Definition	Tips
Online Harms	Detrimental behaviours and practices occurring in the digital sphere that negatively affect individuals and communities.	Understand the types of online harms and their impact on well-being, privacy, and safety. Report harmful content and seek support when needed.
Online Harassment	Unwanted, aggressive behaviour, such as sending threatening or abusive messages, stalking, or spreading false information.	Block and report harassers. Do not engage with abusive messages. Keep evidence of harassment and seek support from trusted individuals or authorities.
Misinformation and Disinformation	Misinformation is false or misleading information shared without harmful intent; disinformation is deliberately false information to deceive and cause public harm.	Verify information from reliable sources before sharing. Be sceptical of sensational headlines. Use fact-checking websites.
Data Protection and Privacy	Measures to safeguard personal information from unauthorised access and misuse.	Use strong, unique passwords for online accounts. Enable two-factor authentication. Be cautious about sharing personal information online. Review privacy settings regularly.
Content Moderation	The process of monitoring and managing user-generated content to ensure compliance with platform rules and regulations.	Understand the platform's community standards. Use reporting mechanisms to flag inappropriate content. Be aware that content moderation can sometimes result in over-censorship.

Algorithmic Bias and Discrimination	Unfair treatment of individuals or groups by algorithms due to biases in data or design	Be critical of the content suggested by algorithms. Seek diverse perspectives. Advocate for transparency and fairness in algorithm design.
Extremism and Polarisation	The amplification of extreme and divisive content can lead to social division and radicalisation.	Avoid echo chambers by following a variety of sources. Report extremist content. Engage in constructive and respectful discussions.
Cybercrime and Fraud	Illegal activities conducted online, such as hacking, phishing, and financial scams.	Be cautious of suspicious links and emails. Use reputable security software. Monitor your financial accounts regularly for unauthorised transactions.
Psychological and Social Effects	The impact of online harms on mental health and social relationships, such as anxiety, depression, and isolation.	Take breaks from social media. Engage in offline activities. Seek professional help if experiencing severe distress.
WhatsApp	A messaging app used for personal and group communication.	Secure your account with two-step verification. Be cautious of misinformation spread through group chats. Adjust privacy settings to control who can see your profile information.
X (formerly Twitter)	A microblogging platform where users share short messages (tweets).	Use reporting tools to handle abuse. Verify information before sharing. Follow accounts that promote positive and reliable content.

Instagram	A photo and video-sharing platform.	Follow positive accounts. Report harmful content. Be mindful of the potential impact of social media on body image and mental health. Adjust privacy settings to control who can see your posts and stories.
Facebook	A social networking site for sharing updates, photos, and links.	Regularly review and adjust privacy settings. Be cautious about accepting friend requests from strangers. Verify information before sharing.
TikTok	A video-sharing platform known for viral trends.	Use privacy features to control who can see your videos. Be cautious about participating in trends. Report harmful content.
Emoticons/Emojis	Small digital images or icons are used to express emotions, and ideas, or convey messages in text form.	Use emojis thoughtfully; they can help express emotion but may also be misinterpreted. Be mindful of their impact on the tone of your messages. Avoid overuse in formal communication.
Memes	Humorous or cultural images, videos, or text that spread rapidly across the internet, often carrying a particular social or political message.	Memes can entertain, but also be used to reinforce stereotypes or spread misinformation. Be critical of the content shared. Consider whether a meme might be offensive or harmful to others.

Retweeting/Reposting	Sharing someone else's post or tweet with your own followers on platforms like X (formerly Twitter), amplifies the original message.	Be cautious when retweeting—ensure that the content aligns with your values. Retweeting without understanding the context can unintentionally spread harmful content.
Reports	Reporting is the action of notifying a platform about inappropriate content or behaviour that violates its guidelines or terms of service.	Use the report feature when you encounter harassment, fake news, or harmful content. Be sure to include any relevant information or evidence.
Stories	Short, temporary posts on social media platforms (like Instagram or Facebook) that disappear after 24 hours.	Use stories for personal, timely updates. Be cautious about sharing sensitive information, as stories may be viewed by a broad audience. Review who can view your stories in the privacy settings.
Reels	Short, engaging videos on platforms like Instagram and Facebook, designed to be fun and shareable.	Create and enjoy reels to express creativity. Be mindful of the content you post—ensure it aligns with your values and doesn't harm others. Consider privacy settings when sharing personal content in reels.